The Development and Testing of a New Version of the Cognitive Reflection Test Applying Item Response Theory (IRT)

The development and testing of a new version of the Cognitive Reflection Test applying Item Response Theory (IRT)

Caterina Primi (primi@unifi.it)[1]

Kinga Morsanyi (k.morsanyi@qub.ac.uk)[2]

Francesca Chiesi (francesca.chiesi@unifi.it)[1]

Maria Anna Donati (mariaanna.donati@unifi.it)[1]

& Jayne Hamilton (jhamilton40@qub.ac.uk)[2]

[1]Department of NEUROFARBA – Section of Psychology, University of Florence

Via di S.Salvi 12- Padiglione 26 – 50135 Florence (Italy)

[2]School of Psychology, Queen's University Belfast,

Belfast, BT7 1NN, Northern Ireland, UK

Acknowledgments

This project was supported by a British Academy/ Leverhulme Small Research Grant to K. M. and C. P. (Grant reference number: SG 120948)

Abstract

The Cognitive Reflection Test (CRT) is a short measure of a person's ability to resist intuitive response tendencies, and to produce a normatively correct response which is based on effortful reasoning. Although the CRT is a very popular measure, its psychometric properties have not been extensively investigated. A major limitation of the CRT is the difficulty of the items, which can lead to floor effects in populations other than highly educated adults. The present study aimed at investigating the psychometric properties of the CRT applying Item Response Theory (IRT) analyses (a 2-parameter logistic model) and at developing a new version of the scale (the CRT-Long), which is appropriate for participants with both lower and higher levels of cognitive reflection. The results demonstrated the good psychometric properties of the original, as well as the new scale. The validity of the new scale was also assessed by measuring correlations with various indicators of intelligence, numeracy, reasoning and decision-making skills, and thinking dispositions. Moreover, we present evidence for the suitability of the new scale to be used with developmental samples. Finally, by comparing the performance of adolescents and young adults on the CRT and CRT-Long, we report the first investigation into the development of cognitive reflection.

Keywords: cognitive reflection; decision making; heuristics; item response theory; individual differences; test information function; validity.

Introduction

The Cognitive Reflection Test (CRT; Frederick, 2005) is a short test measuring a person's tendency to override an intuitively compelling response, and to engage in further reflection which can lead to a correct solution. As an example, consider the following item: *A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost? _____ cents.* Although the correct response is 5 cents, many participants give the response "10 cents", which seems to pop into mind effortlessly. Indeed, a remarkable property of the CRT is that for each item, almost all participants produce either the normatively correct response, or a typical incorrect (i.e., heuristic) response. That is, the reasoning errors that people make are very systematic. It has been proposed that because the heuristic response comes very quickly and easily (i.e., fluently) to mind, people will be highly confident that this answer is correct, and will be reluctant to revise it (cf. Thompson & Morsanyi, 2012). Indeed, De Neys, Rossi and Houdé (2013) found that people who gave the incorrect heuristic response to the bat and ball problem were 83% confident that their response was correct. Although this was significantly lower than the 93% confidence level reported by the participants who gave the correct response, this still demonstrates the attractiveness of the heuristic response, and that generating this response is accompanied by a high level of confidence. Thus, to be able to produce a correct response, participants have to display an outstanding ability to effectively monitor and correct their impulsive response tendencies (cf., Frederick, 2005). As a result, it is only a small minority of participants who tend to produce correct responses to the tasks.

Frederick (2005, p.26) described the CRT as a "simple measure of one type of cognitive ability", and reported reliable relationships between cognitive reflection and measures of intelligence, as well as decision-making skills. Specifically, cognitive reflection was found to be negatively related to temporal discounting (i.e., the tendency to prefer smaller, immediately available rewards, as compared to larger rewards which will be available later), and positively related to choosing gambles

with higher expected values (Frederick, 2005). Further studies showed that the CRT was also related to some typical heuristics and biases tasks, as well as to logical reasoning ability (e.g., Liberali, Reyna, Furlan et al., 2012; Toplak, West & Stanovich, 2011, 2014). Furthermore, although the CRT correlates with measures of intelligence and numeracy (e.g., Frederick, 2005) it was found to explain additional variance in reasoning and decision-making tasks when it was administered together with measures of intelligence and numeracy (Liberali et al., 2012; Toplak et al., 2011). In line with these findings, Campitelli and Gerrans (2013) demonstrated by means of mathematical modelling that the CRT was not just a measure of numeracy. Other studies showed an association between the CRT and metacognitive skills (Mata, Ferreira & Sherman, 2013), and people's motivation to fully understand causal mechanisms (Fernbach, Sloman, Louis & Shube, 2013), and a negative association between the CRT and superstitious and paranormal beliefs (Pennycook, Cheyne, Seli, Koehler & Fugelsang, 2012). Overall, these results demonstrate that the CRT is a very powerful predictor of a person's ability to make unbiased judgments and rational decisions in a wide variety of contexts.

The aim of the present studies

Despite the widespread use of the CRT, its psychometric properties have not been extensively investigated. Most importantly, there are no available data regarding the dimensionality of the scale. In fact, besides its validity, the only property of the scale which has been investigated was its reliability. In his original publication, Frederick (2005) did not report the reliability of the scale, and most researchers who used the scale followed the same practice. A few exceptions include Liberali et al. (2012), who reported a Cronbach's alpha of .74 in Study 1 and .64 in Study 2, Weller, Dieckmann, Tusler et al. (2012) and Campitelli and Gerrans (2013) who reported respectively .60 and .66, and Morsanyi, Busdraghi and Primi (2014), who reported a Cronbach's alpha of .57 in Experiment 1, and .68 in Experiment 2. Thus, one aim of the present investigation was to address these issues.

Another aim was to develop some new items, and a longer version of the scale. The most important reason for developing a longer version of the scale is related to the difficulty of the original items. Indeed, in his original publication, Frederick (2005) reported that in some university student samples, more than 50% of the respondents scored 0 on the test. Thus, the test might not be suitable for lower ability or less educated participants. This obviously limits the generalizability of the findings and makes the scale unsuitable for populations with lower levels of cognitive reflection (e.g., developmental samples). Additionally, even in the case of highly educated adults, using more items can make it possible to discriminate more precisely between respondents with different levels of cognitive reflection. Indeed, by adding some new items to the scale, it is more likely that a normal distribution of scores can be obtained, instead of the typical finding of a highly skewed distribution (i.e., a high proportion of participants scoring zero on the scale). Finally, given the huge popularity of the scale, some participants might already be familiar with the original items, which obviously weakens the suitability of the original scale for measuring cognitive reflection. Given that the new items are unknown, it is obviously not possible for participants to retrieve the correct responses from memory.

Very recently, Toplak et al. (2014) also developed a longer version of the scale. This scale offers the advantage of presenting four new items, and the new scale (including the original and the new items) had good reliability (a Cronbach's alpha of .74). The new items also showed similar relationships with various measures of rational thinking, reasoning and thinking dispositions as the original scale. Nevertheless, this scale has its own limitations. Most importantly, one of the four new items was not open-ended, but participants had to choose from three response options, which made it possible to generate a correct response by chance. Moreover, for one item (the "barrel problem") only 54% of the participants in Toplak et al.'s sample generated either the heuristic or the correct response. Indeed, the heuristic response was only given by 31% of responders, whereas 46% of responders

generated a different incorrect response[1]. This was in sharp contract with the findings regarding the original items where over 80% of the responders generated either the typical heuristic or the correct response for each item, and 60% or more respondents generated the heuristic response. Additionally, the dimensionality of the scale was not analyzed, it is unclear how the most suitable items to be included in the scale were identified (for example, it is not stated whether other items were considered in the study, but then discarded), and the results were only based on a single sample of university students (n=160).

The Item Response Theory (IRT) model

We addressed the above issues regarding the existing versions of the CRT by testing the psychometric properties of the original scale and developing a new version, the Cognitive Reflection Test – Long, CRT-Long) through applying Item Response Theory (IRT). We chose to employ IRT analyses, as the application of IRT have potential benefits in testing and improving the accuracy of assessment instruments. More specifically, IRT is a model that provides a linkage between item responses and the latent characteristic assessed by a scale. IRT assumes that each examinee responding to a test item possesses some amount of the underlying ability. At each ability level, there will be a certain probability $P(\theta)$ that an examinee will give a correct response to the item. If one plotted $P(\theta)$ as a function of ability, the result would be a smooth S-shaped curve (see Figure 1). This S-shaped curve, known as the Item Characteristic Curve (ICC), describes the probability of a correct response to an item as a function of the possessed ability. This probability will be small for examinees of low ability and large for examinees of high ability. The probability of a correct response is near zero at the lowest levels of ability. It continues to increase up to the highest levels of ability, where the probability of

[1] During our scale construction process (see the description in the method section of Study 2) we also considered the "barrel problem", but discarded it for similar reasons (i.e., most participants who gave an incorrect response did not generate the supposed heuristic response).

producing a correct response approaches 1. Each item in a test will have its own item characteristic curve depending on its specific properties.

——

Insert Figure 1 around here

——

Although a number of different IRT models exist, the most commonly employed one is the two-parameter logistic model (2PL) which assumes a single underlying ability and two item parameters: the difficulty parameter (*b*) and the discrimination parameter (*a*). Under IRT, the difficulty of an item describes where the item functions along the trait, and it can be interpreted as a location index with regard to the trait being measured. For example, a less difficult item functions among the low-trait respondents and a more difficult item functions among the high-trait respondents. The second item property is discrimination, which describes how well an item can differentiate between examinees with different levels of ability. The slope corresponds to item discrimination. It describes how rapidly the probabilities change in correspondence with changes in ability levels. This property is essentially reflected by the steepness of the item characteristic curve. The steeper the curve, the better the item can discriminate between levels of ability. The flatter the curve, the less the item is able to discriminate.

Additionally, IRT makes it possible to assess the measurement precision of the test through the *Test Information Function* (TIF) which, instead of providing a single value (e.g., coefficient alpha) for reliability, evaluates the precision of the test at different levels of the measured construct (see Embreston & Reise, 2000; Hambleton, Swaminathan, & Rogers, 1991). The information function is the expected value of the inverse of the error variances for each estimated value of the underlying construct [$I(\theta) \approx 1/SE^2(\theta)$]. The associated reliability is 1 minus the inverse of the information the test provides [*r*

= 1- (1/I)]. This means that the more information the test provides at a particular ability level, the smaller the error associated with ability estimation is, and the higher the test's reliability. In terms of graphical presentation, the Test Information curve shows how well the construct is measured at different levels of the underlying construct continuum (e.g., a peak means that the scale is reliable in a narrow region of the latent distribution).

Summary

In sum, in the present work we analyzed the properties of the original CRT items defining their precision and difficulty level using IRT (Study 1). In Study 2 we developed a new scale (CRT-L), which could be used as an extended version of the original one in order the overcome its limitations, in particular, that the original items might be familiar to respondents, and that they are only appropriate for highly educated adult samples. We investigated the validity of both the original CRT, and the CRT-L, by measuring their correlations with intelligence, decision-making and reasoning skills, thinking dispositions, and measures of numeracy. We aimed at demonstrating that the new items and the CRT-L would show similar correlations with all of these measures as the original CRT. Finally, to test the suitability of the scale to be used with younger and less educated respondents, we investigated the performance of secondary school students on the CRT-L(Study 3). This study also presents a comparison between adolescents and young adults with regard to their tendency to produce heuristic and correct responses to the CRT items.

Study 1

In Study 1 we used IRT analyses to test the psychometric properties of the original CRT.

Methods

Participants

The participants were 438 university students (Mean age = 20.4 years; *SD* = 4.1; 75% female; 26% Italian and 74% British) attending the first or second year of university at the School of

Psychology and the School of Medicine in Florence (Italy) and Belfast (United Kingdom). All students participated on a voluntary basis.

Materials and procedure

The *Cognitive Reflection Test* (CRT**)** is composed of three items[2] (Frederick, 2005). The items are open-ended and there is no time limit to solve them. The responses were classified as correct, heuristic (i.e., the typical intuitive response that quickly comes to mind) or atypical (i.e., neither heuristic, nor correct). A single composite score was computed based on the sum of correct responses. Each participant individually completed the CRT in a self-administered format. Students were instructed to take as much time as they needed to complete the tasks. The average administration time was about 3 minutes. Cronbach's alpha for the current sample was .65.

Results

Item descriptives were calculated and, as expected, the vast majority of participants generated either the correct or the heuristic response for each item (see Table 1).

Then, using the correct responses, the factorial structure of the original CRT was tested evaluating the presence of local dependence (LD) that is a term used to describe excess covaration among item responses that is not accounted for by a unidimensional IRT model. To investigate LD, the χ^2 LD statistic (Chen & Thissen, 1997) was used. A value of 10 or greater is considered noteworthy. Results showed that a single factor model adequately represented the structure of the CRT given that none of the LD statistics were greater than .7. The factor loadings were all significant ($p < .001$) and high in value[3] (see Table 1).

[2] The Italian version of the CRT was obtained using a forward-translation method: two non-professional translators worked independently, and then they compared their translations with the purpose of assessing equivalence, taking into account the differences in cultural context. Subsequently, a group of five people read this first version, revised it, and eventually obtained a final form which can be considered consistent with the original one . Please email the Authors for a copy of the Italian version of the scale.

[3] The same analyses were conducted separately on the Italian and English samples to check for equivalence. The results attested that the Italian and English versions of the scale shared the same one-factor structure, and similar patterns of factor loadings. This made it possible to merge the data across the British and Italian samples to perform IRT analyses.

——

Insert Table 1 around here

——

Having verified the assumption that a single continuous construct accounted for the covariation between item responses, unidimensional IRT analyses were performed. The two-parameter logistic model (2PL) was tested in order to estimate the item difficulty and discrimination parameters. The parameters were estimated by employing the marginal maximum likelihood (MML) estimation method with the EM algorithm (Bock & Aitkin, 1981) implemented in IRTPRO software (Cai, Thissen, & du Toit, 2011). In order to test the adequacy of the model, the fit of each item under the 2PL model was tested computing the *S-χ2* statistics. Each item had a non-significant S-χ^2 value, indicating that all items fit under the 2PL model. Concerning the difficulty parameters (*b*), the results showed that the parameters ranged from .08 ± .07 to .70 ± .12 logit across the continuum of the latent trait. The logit is the logarithm of the *odd*, that is, the ratio between the probability of producing a correct response and the probability of responding incorrectly. Given that the difficulty parameter *b* has a mean of zero and SD of 1.0, all the CRT items had an above-medium level of difficulty with item 3 being easier than the other items. With regard to the discrimination parameters (*a*), following Baker's (2001) criteria, all three items showed large discrimination levels (*a* values over 1.34). Figure 2 shows the item characteristics curves used in IRT. All items were located in the positive range of the trait, indicating the regions where they function better. Discrimination is represented by the steepness of the curve. All items showed a high slope, indicating their ability to distinguish between respondents with different levels of the trait around their location.

——

Insert Figure 2 around here

——

Finally, in order to identify the level of ability that is accurately assessed by the scale, the Test Information Function (TIF) was analyzed. The peak of the TIF is where measurement precision is the greatest. The TIF estimated under the 2PL model showed that the original CRT scale was sufficiently informative for the middle level of the trait within the range of trait from -0.4 to 0.4 standard deviations around the mean (fixed by default to 0; see Figure 3). Specifically, the overall distribution of information reveals that the CRT provides more information around the mean level of the trait (I = 6.10, r = .84) and that information is relatively consistent between trait levels -.4 (I = 3.61, r = .72) and .4 (I = 5.02, r = .80), but information declines sharply below 3 (corresponding to a reliability level < .70) in the rest of the trait range. Accordingly, the three items of the CRT seem incapable of differentiating low from lower medium as well as higher medium from high levels of the latent trait.

——

Insert Figure 3 around here

——

Discussion

Despite the widespread use of the CRT scale, its dimensionality has never been tested. The present study established the unidimensionality of the CRT , which is a desirable characteristic, as it makes it suitable to be used as a population screen of the cognitive reflection trait.

Applying IRT analyses for the first time to measure the psychometric properties of the CRT, we also demonstrated that the original CRT items had high discriminative power(i.e. they were able to

distinguish between respondents with different levels of the cognitive reflection trait), and they had a medium level of difficulty.

Study 2

This study was aimed at developing a longer versionof the CRT scale, with the addition of new items which have similar properties to the original ones, but possess different difficulty parameters. In this way, it would be possible to obtain a scale which could also be used to measure the cognitive reflection ability trait in lower ability or less educated participants. Similarly to Experiment 1, we conducted IRT analyses to estimate item parameters. Additionally, the validity of the new items, and of the CRT-Long (CRT-L) were studied. In particular, we aimed at demonstrating that the CRT-L showed similar correlations with various measures of intelligence, numeracy, decision-making skills, reasoning ability, and thinking dispositions as the original CRT, and confirming that the CRT-L is a better predictor of decision making performance than numeracy and fluid intelligence. This is important, as previous research showed that the CRT was more than just a test of numeracy and intelligence (see Liberali et al., 2012; Toplak et al., 2011). Moreover, we measured gender differences, and in line with the original study by Frederick (2005), we expected to find that males would perform better on the CRT-L. We also repeated the IRT analyses of the original CRT items in order to demonstrate the robustness of the results across different samples.

Methods

Participants

The participants were 988 students (Mean age = 20.2 years; *SD* = 1.8; 63% female; 76% Italian and 24% British) attending the senior year of high school (40%) or undergraduate university courses(60%) at the School of Psychology and the School of Medicine in Florence (Italy) and Belfast (United Kingdom). All students participated on a voluntary basis, and the university students received ungraded course credit for their participation.

Materials

The *Cognitive Reflection Test - Long* (CRT-L)[4] was composed of the three original items (Frederick, 2005) and three new items (see Appendix A).

The development of the new items followed three steps in which some items were eliminated and others modified (see Appendix B for the list of items considered). In the first step, we piloted 6 new items on 52 Italian university students. This version included 5 new items developed by Shane Frederick (personal communication, July 2012), and an additional item based on Van Dooren, De Bock, Hessels, Janssens and Verschaffel (2005). The initial results showed that from the 6 new items, only a single item (item 5 in Appendix B, which is also included in Toplak et al.'s long version of the CRT) displayed similar properties to the original CRT items, in that most participants either produced the correct response or a typical incorrect response.

Based on these initial findings, we kept one item unchanged (item 5 in Appendix B), we modified the numbers in three items (items 1, 4 and 6), discarded two items, and generated 3 new items (items 7-9) where we tried to reproduce the structure of the original CRT items, but we used different content and numbers. The selection and modification of the items was performed with the aim of identifying items with similar, but also with lower and higher levels of difficulty than the three original ones. We administered the original and the new items to a sample of 58 Italian university students. The results showed that after the modifications, the four new items included in our first version of the scale, and one item (item 9) that we added later displayed the desired characteristics (i.e., that most participants either produced the heuristic or the correct response). In sum, this version of the scale included five new items: three items developed by Frederick, from these two were modified (i.e., the numbers included in them were changed), a modified version of Van Dooren et al.'s (2005) problem, and an additional item developed by us. This version was administered to a sample of 158 Italian

[4] The Italian version of the CRT-L was obtained using a forward-translation method, using the same procedure as for the original CRT scale.

students and applying IRT two other items were discarded. One item (item 6) was discarded because it had a lower level of discrimination than the other items, and another one (item 1) had a bad fit under the 2PL model once item 6 was removed. In sum, the final version of the CRT-L was composed of six items: three original items (Frederick, 2005) and three new items (see Appendix A for the final version of the scale). A single composite score was computed based on the sum of correct responses. Coefficient alpha for the current sample was .76.

Measure of intelligence

Set I of the *Advanced Progressive Matrices* (APM-Set I; Raven, 1962) is a measure of fluid intelligence, and it was used as a short form of the Raven's *Standard Progressive Matrices* (SPM, Raven, 1941). Set I of the APM is composed of 12 items that increase in their level of difficulty, and cover the full range of difficulty of the SPM (Raven, 1962). These items consist of a series of perceptual analytic reasoning problems, each presented in the form of a matrix. In each case, the lower right corner of the matrix is missing and the participant's task is to determine which of eight possible options fits into the missing space such that the row and column transformations are satisfied. Using IRT analysis procedures, the short form of the SPM was found to have high reliability and validity (Chiesi, Ciancaleoni, Galli, & Primi, 2012).

Measures of mathematical and probabilistic reasoning

The *Probabilistic Reasoning Questionnaire* (PRQ, Primi, Morsanyi & Chiesi, 2014) was designed to measure proportional reasoning and basic probabilistic reasoning skills. The scale consists of 16 multiple-choice probabilistic reasoning questions. The items include questions about simple, conditional, and conjunct probabilities, and the numerical data are presented in frequencies or percentages. A single composite score was computed based on the sum of correct responses. Coefficient alpha for the current sample was .68.

The *Numeracy Scale* (NS, Lipkus, Samsa & Rimer, 2001) was composed of 11 items that assess basic probability and mathematical concepts including simple mathematical operations on risk magnitudes using percentages and proportions. A single composite score was computed based on the sum of correct responses. Coefficient alpha for the current sample was .59.

The *Math Fluency* subtest of the Woodcock-Johnson III Tests of Achievement (MF, Woodcock, McGrew, & Mather, 2001) was used as a measure of arithmetic skills. This test assesses the ability to solve simple addition, subtraction, and multiplication problems quickly. Participants were asked to work through a series of problems as quickly and accurately as possible within a three-minute time limit. The total number of items that were correctly solved was calculated to provide a math fluency score.

The *Subjective Numeracy Scale* (SNS, Fagerlin, Zikmund-Fisher, Ubel, Jankovic, Derry & Smith, 2007) is a subjective measure (i.e., self-assessment) of quantitative ability which was developed with the aim of distinguishing between low- and high-numerate individuals, but in a less aversive and quicker way than it is possible with objective tests of numeracy. An example item is: "How good are you at working with fractions?" The items have to be rated on a 6-point Likert scale, which are labelled differently, depending on the question asked (e.g., ranging from *1=not good at all* to *6=extremely good; or* 1=*never* to *6=very often*). A single composite score was computed based on participants' ratings of each item. Coefficient alpha for the current sample was .62.

Measures of reasoning abilities

In the *Conditional Probability task* (based on Gauffroy & Barrouillet, 2009) participants were presented with a visual display of 8 cards, and they were given the following information. "There are 8 cards in a pack. All the questions concern these 8 cards." Some of the cards presented to participants were black, and some were white, and some of them had a circle printed on them, others had a square printed on them. The participants were then asked 4 questions about the cards, which had the

following format: “How likely are the following statements to be TRUE/FALSE of a card drawn randomly from the pack? If the card is black then there’s a square printed on it.” Participants had to provide a response in the format of: “__ out of __” (i.e., participants had to fill the gaps with the relevant numbers). Two questions asked about the likelihood of the statement being true, and two questions asked about the likelihood of the statements being false. Performance on the task measures hypothetical thinking skills, and it was found to be related to participants’ general intelligence (Evans, Handley, Neilens & Over, 2007). In the present sample a single composite score was computed, based on the overall number of correct responses. Cronbach’s alpha for the task in the current sample was .82.

The *Conditional Reasoning Task* consisted of 12 problems. Four problems were included which corresponded to each of the following inference forms: modus ponens (MP), denial of the antecedent (DA), and affirmation of the consequent (AC). All the problems had familiar, everyday content (e.g., “If the radio is turned on, then you will hear music. The radio is not turned on. Is it necessary that: You will not hear music?”) Participants had to respond by choosing either “yes” or “no, it’s not necessary”. The participants were provided with detailed instructions, and were asked to imagine that the first two statements were always true, before deciding if the conclusion necessarily followed. Before completing the task, the participants were also presented with a practice problem to familiarize them with the presentation format. A single composite score, based on the total number of correct responses was computed. Cronbach’s alpha in the current sample was .65.

The *Transitive Inference* task included 12 problems (which were closely modelled on Morsanyi, Devine, Nobes & Szucs, 2013). Four of the problems had believable conclusions, four had unbelievable conclusions and the remaining four had belief-neutral conclusions. In addition to this, half of the problems were valid (i.e., the conclusion followed from the premises) and half were invalid. Valid problems followed one of the following structures: *A>B and B>C, is A>C?* or *A>B and C>A, is C>B?* Invalid problems followed one of these structures*: A>B and B>C, is C>A?* or *A>B and C>A, is B>C?*

In the case of belief-laden problems, half of the problems were conflict problems (i.e., where the believability and validity of the problem cued a different response), and half of them were non-conflict problems. Problems from the various categories were mixed together, and were presented in the same order to all participants. Detailed instructions accompanied the problems, asking participants to accept the first two statements to be true, even if they were not true in real life. Participants were then asked to determine if the third statement logically followed from the first two statements, or if it did not necessarily follow. The problems had the following format. "Imagine that the following is always true: Sharks are more dangerous than cuttlefish. Now imagine that this is also true: Cuttlefish are more dangerous than sea urchins. Is it necessary that: Sharks are more dangerous than sea urchins?" Participants had to respond by choosing either "yes" or "no, not necessarily". A single composite score, based on the total number of correct responses was computed. Cronbach's alpha in the current sample was .75.

Measures of decision-making skills

The *Risk Seeking Behaviour questionnaire* was composed of 8 items adapted from Frederick (2005). For each item participants indicated their preference between a certain, smaller gain and some probability of a larger gain. A composite score was created by summing the responses that indicated the preference to take a risk. A higher score indicated a higher preference to take a risk with the hope of obtaining a larger amount of money. Cronbach's alpha for this scale was .57.

The *Intertemporal Behaviour questionnaire* was composed of 5 items adapted from Frederick (2005). For each item, participants indicated their preference for a smaller amount of money now or a larger amount of money later. A composite score was created by summing the responses that indicated a preference to wait in order to obtain a larger amount of money. A higher score indicated a higher preference to wait for obtaining a larger amount of money. Cronbach's alpha for this scale was .47. Although the reliability of this scale was lower than .5, which is considered to be a satisfactory level of

reliability (cf. Rust & Golombok, 1999), given that a very similar measure was used by Frederick (2005), we report the results regarding this measure as well.

In the *Framing Task* participants chose between riskless and risky alternatives which had identical expected values, under both a gain-framing and a loss-framing condition. This task was an adaptation of Tversky and Kaheneman's (1981) famous Asian disease problem. The order in which participants were presented with each framing condition was counterbalanced, and each participant was exposed to both versions. The following scale was used to indicate preference: 1.Strongly favor option A; 2.Favour option A; 3.Slightly favor option A; 4.Slightly favor option B; 5. Favor option B; 6 Strongly favor option B. The problem was scored by subtracting the positive frame ratings from the corresponding negative frame ratings. A difference score of 0 indicated the absence of a framing effect and a positive score different form zero represented the presence of a framing effect. Respondents were classified by creating two groups: resistant to framing (difference score of 0) and susceptible to framing (a positive difference score different from zero).

Measure of thinking dispositions

The *Superstitious Thinking Scale* (STS, Kokis, MacPherson, Toplak, West, & Stanovich, 2002; Italian version: Chiesi, Donati, Papi, & Primi, 2010) was composed of eight Likert-type items using a 5-point scale ranging from totally false to totally true, yielding a maximum score of 40. Higher scores represent high levels of superstitious thinking. Coefficient alpha for the current sample was .80.

Procedure

The participants individually completed the measures in a self-administered format in the classroom. Each task was briefly introduced, and instructions for completion were given. The answers were collected in a paper-and-pencil format. All participants completed the CRT-L, whereas the other measures were administered as follows. All Italian participants were administered the framing task, and one sample (N = 201) was administered the APM, the NS, and the risk seeking behaviour measure, and

another sample (N = 393) worked through the STS and the intertemporal behavior questionnaire. A British subsample (N = 234) was administered the PRQ, the SNS and the conditional probability tasks. Another British subsample (N = 59) was administered the math fluency test, the conditional reasoning test, and the transitive inference test. Administration time ranged from half an hour to an hour and a half.

Results

The item properties of the CRT scale

Item descriptives were calculated and, in line with the results of Study 1 the vast majority of participants generated either the correct or the heuristic response for each item (see Table 2).

Then, using the correct responses, we tested the unidimensionality of the scale, evaluating the presence of local dependence (LD). The results confirmed that a single factor model adequately represented the structure of the CRT since none of the LD statistics were greater than .7. Factor loadings were all significant ($p < .001$) and high in value[5] (see Table 2).

Having verified the assumption that a single continuous construct accounted for the covariation between item responses, unidimensional IRT analyses were performed. In order to test the adequacy of the model, the fit of each item under the 2PL model was tested computing the $S\text{-}\chi 2$ statistics. Each item had a non-significant S-$\chi 2$ value, indicating that all items fit under the 2PL model. Concerning the difficulty parameters (b), the results showed that the parameters ranged from -.12 ± .07 to .43 ± .07 logit across the continuum of the latent trait.

As in Study 1 item 3 was easier than the other two items which had above-medium levels of difficulty. Concerning the discrimination parameters (a), following Baker's criteria (2001), all three items showed large discrimination levels (a values over 1.34; seeTable 2). As in Study 1 all items were located in the positive range of the trait, indicating the regions where they function better.

[5] As in Study 1, we confirmed that the Italian and English versions of the scale shared the same one-factor structure, and similar patterns of factor loadings.

Discrimination is represented by the steepness of the curve. The steeper the curve, the better the item can discriminate. All items showed a high slope indicating their ability to distinguish between respondents with different levels of the cognitive reflection trait around their location.

In sum, the comparison between Studies 1 and 2 confirmed the consistency of the properties of the CRT items across samples. the results showed that the items were able to distinguish between respondents with different levels of the cognitive reflection trait around their location which was above the mean.

The item properties of the CRT-L

As a preliminary step, item descriptives were calculated for the CRT-L. We confirmed that the participants mostly generated either the correct or the heuristic response for each item (see Table 3).

Two new items (items 5 and 6) showed similar percentages of correct responses as the original items, whereas item 4 was easier than the other items. Additionally, whereas a high proportion (30%) of participants scored zero on the original scale, on the CRT-L only 9% of participants scored zero.

——

Insert Table 2 around here

——

Preliminarily, we tested the assumption of unidimensionality with the standardized local dependence LD[6]. None of the standardized χ^2 indices of LD approached the value of 10. Factor loadings were all significant ($p < .001$), ranging from .70 to .85 (see Table 2). Having verified the assumption that a single continuous construct accounted for the covariation between correct responses to each item, unidimensional IRT analyses were performed.

[6] The Italian and English versions of the scale shared the same one-factor structure, and similar patterns of factor loadings. This made it possible to merge the data across the British and Italian samples to perform IRT analyses on the CRT-L.

In order to test the adequacy of the model, the fit of each item under the 2PL model was tested computing the *S-χ2* statistics. Each item had a non-significant S- *χ2* value (see Table 2), indicating that all items fit under the 2PL model. Concerning the difficulty parameters (*b*), the results showed that the parameters ranged from -.1.19 ± .09 to 0.42 ± .06 logit across the continuum of the latent trait (see Table 2). Concerning the discrimination parameters, all new items showed large discrimination levels (*a* values over 1.34 (see Table 2). Figure 4 shows the Item Characteristic Curves used in IRT to provide visual information of the item characteristics. Severity is represented by the location of the curve along the trait. Concerning the location of the new items, item 4 was located in the negative range, that is, it functions better in lower levels of the trait, item 5 was located in the positive range of the trait, that is, it is able to measure higher levels of the trait while item 6 had a medium level of difficulty. All items showed a high slope, indicating their ability to distinguish between respondents with different levels of the trait around their location.

——

Insert Figure 4 around here

——

Finally, in order to identify the level of ability that is accurately assessed by the scale, the TIF was analyzed. The test information function displayed in Figure 4 makes it possible to compare the varying measurement precision across the construct continuum for the CRT and the CRT-L. It can be seen that the CRT-L's information curve has higher information values associated with a larger range of values of the construct as compared to the CRT. On the one hand, the overall distribution of information reveals that the CRT provides more information around the mean level of the trait (I = 4.93, r = .80) and that information is relatively consistent between trait levels -0.4 (I = 4.29, r = .77) and 0.4 (I = 3.67, r = .73), but information declines sharply below 3 (corresponding to a reliability level < .70) in the rest of the trait range. Accordingly, the three items of the CRT seem incapable of

differentiating low from lower medium as well as higher medium from high levels of the latent trait. On the other hand, the overall distribution of information of the CTR-L reveals that it provides the most information around the mean level of the trait (I = 6.13, r = .84) but the information provided is relatively consistent across a larger range of the trait from -1.0 (I = 3.74, r = .73) to 1.0 (I = 3.47, r = .71), dropping below 3 under -1.2 and over 1.2. Accordingly, the six items of the CRT-L seem capable of differentiating from lower medium to higher medium levels of the latent trait and, as a result, they allow for a better assessment of individual differences in the cognitive reflection construct than the CRT.

——

Insert Figure 5 around here

——

The validity of the CRT-Long scale

Descriptive statistics for each individual differences measure are reported in Table 4. Concerning the validity measures, Pearson product-moment correlations for the CRT and the CRT-L attested that the relationships which were investigated were significant and in the expected directions (see Table 3). Indeed, not only the CRT-L scale, but also the new items showed similar relationships with each construct which was investigated as the CRT.

——

Insert Table 3 around here

——

Both CRT measures correlated positively with the APM[7], which was in line with previous studies that reported a relationship between the CRT scores and measures of intelligence (Frederick, 2005; Toplak et al., 2011). Concerning numeracy, we obtained a positive correlation with the CRT measures, and the values appeared to be similar to the values reported in previous studies employing the CRT (Cokely & Kelly, 2009; Liberali et al. 2011; Weller et al., 2013). In addition, we also found that the CRT scales showed similar correlations with some further measures of mathematical and probabilistic reasoning skills, including arithmetic skills (i.e., math fluency) and subjective numeracy.

With regard to the decision-making measures, Frederick (2005) observed that the original CRT was positively related to decisions in risky choice tasks. Specifically, higher CRT scores were related to more risky choices than low CRT scores. Our findings are in line with previous results, confirming a positive correlation between risky choice and the CRT measures. We also found a positive correlation with intertemporal behavior (i.e., the tendency to prefer a larger reward that will be available later, over a smaller, immediately available reward), in line with Frederick (2005) and Toplak et al. (2011).

Previous studies reported that the CRT was more than just a measure of intelligence or numeracy, as it predicted additional variance in reasoning and decision-making tasks (see Liberali et al, 2012; Toplak et al., 2011; 2014). One sample of participants completed a test of numeracy and fluid intelligence, together with the CRT-L and the risk seeking behavior scale. Thus, in this sample we were able to test the hypothesis that the CRT-L score was a more powerful predictor of risk seeking behavior than fluid intelligence and numeracy. We conducted a simultaneous multiple regression analysis with risk seeking behavior total score as a dependent variable and CRT-L score, numeracy, and fluid intelligence as predictors. The results confirmed that, when entered simultaneously, the CRT-L score was a significant predictor (see Table 4) of risk seeking behavior whereas, numeracy and intelligence were not.

[7] Regarding intelligence, the normality indices of the APM total score distribution revealed that departures from normality were acceptable (Skewness -.51 and Kurtosis =.34; Marcoulides & Hershberger,1997).

——

Insert Table 4 around here

——

Another sample of participants completed the probabilistic reasoning questionnaire, together with the subjective numeracy scale and the math fluency scale. In this sample we were able to test the hypothesis that the CRT-L might be a better predictor of probabilistic reasoning skills than math fluency and subjective numeracy. When math fluency and subjective numeracy were entered into the regression equation together with the CRT-L, they did not significantly predict probabilistic reasoning performance, whereas the CRT-L was a significant predictor (see Table 5).

——

Insert Table 5 around here

——

Concerning thinking dispositions, we found, as expected, a negative correlation with superstitious thinking.

With regard to our measures of reasoning ability (i.e., conditional probability reasoning, everyday conditional reasoning, and transitive inferences), again the relationships between the CRT scales and the reasoning tasks were similar across the two versions of the CRT, and cognitive reflection was positively related to reasoning ability (see also Campitelli & Gerrans; 2013; Toplak et al., 2011).

Regarding the frame task, we first explored the effect of the order of exposure to the different frames. Finding no order effect, we collapsed the different order conditions in subsequent analyses. A *t* test for independent samples (resistant to framing *vs.* susceptible to framing) was conducted with CRT scores as the dependent variable. With regard to the original CRT scale, respondents who were resistant to framing had higher scores (M = 1.43; SD = 1.2) than respondents who were susceptible to framing effects (M = .82; SD = .99); t (488) = -5.90; $p < .001$; d = .55). Considering only the CRT-L new items

score the difference was also significant (t (487) =- 4.39*; p < .001; d = .41) with respondents who were resistant to framing having higher scores (M = 1.64; SD = 1.01) than respondents who were susceptible to framing (M = 1.25; SD = .91). Finally, considering the CRT-L score the difference was again significant (t (474) = -5.58*; p < .001; d = .53) with respondents who were resistant to framing having higher scores (M = 3.11; SD = 1.9) than respondents who were susceptible to framing (M = 2.13; SD = 1.7). Differently from Toplak et al. (2014) but in line with Frederick (2005), our results attested the relationship between the CRT and susceptibility to the framing effect.

Finally, concerning gender differences, in the original CRT, males (M = 1.63; SD = 1.1) scored significantly higher than females (M = 1.25; SD = 1.1; t (939) = 5.06*; p < .001; d = .34). Considering only the CRT-L new items score the findings were consistent with the original CRT results, confirming a significant gender difference (t(922) = 3.67; p < .001; d = .20), with a better performance in males (M = 1.84; SD = .99) than in females (M = 1.64; SD = .99). The same significant difference was found considering the CRT-L (t(908) = 4.86; p < .001; d = .32), with a better performance in males (M = 3.54; SD = 1.9) than in females (M = 2.91; SD = 1.8).

With regard to the origin of the gender differences, we tested the possibility that these were related to the numerical content of the problems. Indeed, gender differences have been found to affect mathematical problem solving, and the gender gap tends to be greater in highly select samples of young adults (e.g., Hyde, Fennema, & Lamon, 1990). Although the mathematical computations required to solve the CRT problems are simple, women might experience higher levels of anxiety, and might be less confident in their ability to solve these problems (e.g., Beilock, 2008).

An Italian subsample completed the numeracy scale together with the CRT. Our preliminary analyses showed that female students scored significantly lower (M=7.05; SD=1.71) than males (M=7.70; SD=1.38; t(199)=2.46, p=.015, Cohen's d=.35) on the numeracy scale. We run an ANCOVA with gender (male/female) as a within-subjects factor and numeracy as a covariate on the CRT-L

scores. Although the effect of gender was still significant ($F(1,159)=4.81$, $p=.030$, $\eta_p^2=.03$), numeracy was also a significant covariate ($F(1,159)=31.08$, $p<.001$, $\eta_p^2=.17$).

One of the British samples completed the subjective numeracy scale, and the PRQ together with the CRT. We first checked for gender differences on the PRQ and on the subjective numeracy scale. On the PRQ females ($M=14.95$; $SD=1.31$) and males ($M=15.08$; $SD=1.22$) performed equally well ($p=.491$). However, on the subjective numeracy scale females ($M=36.92$; $SD=5.37$) scored significantly lower than males ($M=38.77$; $SD=5.61$; $t(212)=2.38$, $p=.018$, Cohen's $d=.33$). We run an ANCOVA with gender (male/female) as a within-subjects factor and subjective numeracy as a covariate on the CRT-L scores. Subjective numeracy was a significant covariate ($F(1,203)=6.55$, $p=.011$, $\eta_p^2=.03$), and the effect of gender was no longer significant ($p=.102$).

Discussion

In this study we confirmed the psychometric properties of the CRT items using a different sample, in order to demonstrate the robustness of the IRT results. Indeed, just as in Study 1, the items had high discriminative power, and a medium level of difficulty.

Regarding the CRT-Long scale the IRT analyses showed that the three new items had high discriminative power, and their location was more distributed along the ability trait in comparison with the original version with the additional advantage of including new items that participants are unfamiliar with. These analyses reconfirmed the suitability of the original items for measuring the cognitive reflection trait, and also demonstrated that the new 6-item version measures accurately a wider range of the cognitive reflection trait.

Concerning the validity of the new scale, our study is unique in terms of the range of domains covered by our measures. We demonstrated that the CRT-Long showed similar correlations with various measures of numeracy, reasoning and decision making skills, intelligence and thinking dispositions as the original CRT. Indeed, the correlations that we obtained with these measures were

similar across the different versions of the scale, and also similar to the correlations reported in previous studies. We also included a set of novel measures of validity (i.e., probabilistic thinking, conditional probability reasoning, everyday conditional reasoning and transitive inferences), which were related to previously investigated constructs, but also furthered our understanding of the range of thinking skills that are assessed by the CRT. Previous studies (Liberali et al, 2012; Toplak et al., 2011; 2014) demonstrated that the CRT predicted additional variance in reasoning and decision making tasks when it was administered together with measures of intelligence or numeracy. Although some of our measures were only completed by a subsample of our participants, we were able to run two simultaneous regression analyses where we compared the predictive power of the CRT-L with that of measures of numeracy and fluid intelligence. These analyses showed that the CRT-L was the strongest predictor of both probabilistic reasoning skills and risk seeking behavior, which further demonstrates the versatility of the CRT-L.

Several previous studies reported gender differences in cognitive reflection (e.g., Campitelli & Gerrans, 2013; Frederick, 2005; Toplak et al., 2014). We hypothesized that the reason for this might be that the CRT has a significant math component. Indeed, in one of our samples, controlling for numeracy significantly reduced the gender difference in the CRT, and in another sample controlling for subjective numeracy eliminated the gender difference. We will return to this issue in the general discussion.

In summary, although the original CRT has appropriate psychometric properties, the new scale offers high precision across a wider range of the cognitive reflection ability trait, and reduces the proportion of participants who score zero on the scale. The new items also share the essential property of the original items that, although the problems are open-ended, the vast majority of participants produce either the correct or the typical heuristic response to problems. This is an important advantage of the present scale over the version proposed by Toplak et al. (2014). Being able to also measure

precisely lower levels of cognitive reflection opens up the possibility of using the scale with populations other than highly educated adults (such as developmental samples, and older or less educated adults). The aim of Study 3 was to explore the suitability of the new scale to be used in developmental research.

Study 3

The results reported by Frederick (2005) demonstrated the difficulty of the original items. Indeed, in some university samples more than 50% of the participants scored 0 on the test. For this reason the original test might not be suitable for lower ability or less educated participants. In Study 2 we demonstrated using IRT that the items of the CRT-Long displayed high discriminative power and their location was more distributed along the ability trait. For this reason we expected that the new scale would not only be more suitable to discriminate between participants with high levels of the cognitive reflection trait, but it would also be more suitable to measure cognitive reflection in participants with lower levels of the trait, such as adolescents. Indeed, given that even in some adult populations the majority of participants score zero on the original scale, floor effects in developmental populations seem almost inevitable.

Making the scale more suitable for younger participants also means that it is possible to investigate developmental patterns in heuristic and correct responding. Thus, an important aim of this study was to compare the performance of adolescent and adult participants on the scale. Although the development of cognitive reflection has not been investigated before, there are good reasons to expect that adolescents would perform more poorly on the CRT and CRT-L than adults (i.e., they would produce fewer correct responses). Although very few studies (e.g., Chiesi et al., 2011; Klaczynski, 2001a; Morsanyi & Handley, 2013) compared the performance of adolescents and adults on tasks which elicit a conflict between heuristic and normative response tendencies, these studies reported a general increase in normatively correct responses, whereas heuristic responding either decreased or

remained stable across development. Studies which investigated age-related changes during adolescence (e.g., Klaczysnki, 2001a,b; Klaczynski & Cottrell, 2004) also reported similar changes between early and middle or late adolescence.

In terms of establishing developmental patterns in heuristic and normative responding, two important properties of the CRT items deserve special attention. One of these is that whereas the normative standards used to evaluate some typical heuristics and biases tasks have been famously questioned (see e.g., Gigerenzer, 1996; Hertwig, Benz & Krauss, 2008), in the case of the CRT items the normative standards used are indisputable (i.e., the typical heuristic responses are clearly incorrect). Another very important characteristic of these items is that although the heuristic and correct responses are mutually exclusive, given that the problems are open-ended, heuristic and normative responses are not explicitly pitted against each other. This makes it possible to investigate heuristic and normative responding relatively independently (i.e., the ability to resist heuristic response tendencies does not necessarily imply that a participant will produce a correct response).

In summary, the aim of our final study was to investigate the development of cognitive reflection for the first time, including an investigation of the tendencies to produce both heuristic and correct responses.

Methods

Participants

The participants were 70 British adolescents (29 females, between the ages of 11 years 7 months and 14 years 4 months; mean age=13 years) recruited from a secondary school in a small town in Northern Ireland, and 287 young adults (199 females, between the ages of 18 and 25 years; mean age=19 years 2 months) recruited from undergraduate courses at Queen's University Belfast. None of the adult participants were involved in Study 2.

Materials and procedure

The participants individually completed the CRT-Long in a self-administered format in the classroom. The task was briefly introduced, and instructions for completion were given. The answers were collected in a paper-and-pencil format. Students could take as long as they needed to solve the problems. The average completion time was about 5 minutes. Cronbach's alpha for the CRT-Long was .79 in the adolescent and .68 in the adult sample.

Results

First we present the distribution of correct and heuristic scores on the CRT-L for the early adolescents and young adults (see Figure 6). As it is apparent, the correct scores of the adult participants showed a close-to-normal distribution, and only 2% scored zero on the scale, which is a great improvement compared to the original scale where 33% of the adult participants scored zero. In the case of the adolescent sample, 49% scored zero on the scale, as opposed to 67% scoring zero on the original items.

With regard to heuristic responses, in the adolescent sample nobody had a score of zero, and 14% of the sample scored 6 (i.e., they gave a heuristic response to every item). In the adult sample 16% scored zero, and 0.3% scored 6. Cronbach's alpha for the scale using the heuristic scores was .61 for adolescents and .66 for young adults. Given these results, it could be advantageous to use the heuristic scores instead of the correct scores in the case of participants with lower levels of cognitive reflection.

——

Insert Figure 6 around here

——

Next we investigated the developmental changes in both heuristic and correct responding (see Table 5). In the case of each item the large majority (78% or more) of both adolescents and adults produced either the correct or the heuristic response. Chi-square tests indicated that young adults were

significantly more likely to respond correctly on each item (X^2>11, p<.001) with the exception of item 2 (p=.56). We also compared the proportion of heuristic responses across groups using chi-square tests. In general, adolescents were more likely to produce heuristic responses than adults. For each item, the chi-squared values were above 22 (p<.001), with the exception of items 2 and 5. For item 5 the chi-squared test still indicated that the younger participants gave more heuristic responses than the older participants (X^2 (1, N = 354) = 3.70, p=.054). In the case of item 2 there was no developmental change in the proportion of heuristic responses (p=.96). In sum, the developmental comparisons showed that in general correct responding increased with age, whereas heuristic responding decreased. The only exception was item 2 where the proportion of heuristic and correct responses remained stable with development.

——

Insert Table 6 around here

——

We also conducted a 2x2 ANOVA on the number of correct responses with item type (original items/new items) as a within-subjects and age group (adolescents/young adults) as a between-subjects factor. The aim of this analysis was to investigate how well the original and the new items discriminated between the two age groups. This analysis is also relevant for judging the suitability of the original and the new items in discriminating between participants with lower and higher levels of the cognitive reflection trait. The ANOVA showed a main effect of item type (F(1,336)=38.19, p<.001, η_p^2=.10), a main effect of age group (F(1,336)=420.71, p<.001, η_p^2=.56), and an item type by age group interaction (F(1,336)=20.00, p<.001, η_p^2=.06). These results indicated that the new items (M=1.76 SD=.97) were easier than the original ones (M=1.17 SD=1.17), young adults (M=3.34 SD1.69) performed better than adolescents (M=1.34 SD=1.73), and the new items discriminated better between

the two age groups (the group difference on these items was 1.30; Cohen's *d*=1.53) than the original items (the group difference on these items was .69; Cohen's *d*=.64).

Discussion

In Study 3 the CRT–L was administered to a sample of adolescents, as well as to a new sample of young adults. The results showed that the CRT-L was a reliable measure of cognitive reflection in both populations. This was the case for both the heuristic and the correct responses. Additionally, a large majority of participants in both age groups (i.e., 78% or more) produced either the correct or the heuristic response for each item.

By adding the new items, we created a scale where the distribution of scores in the case of young adults was close to normal, and only a very small proportion (i.e., 2%) of adult participants scored zero on the test. This is a very important improvement as compared to the original scale, where 33% of our sample scored zero, making it impossible to precisely discriminate between respondents with low levels of the cognitive reflection trait. That is, in a population with low levels of the cognitive reflection trait, the predictive value of the original CRT would be diminished.

With regard to early adolescents, although a large proportion (i.e., 49%) of this population scored zero on the CRT-L, the distribution of heuristic scores was close to normal. Additionally, the scale had good reliability when heuristic scores were used, which raises the possibility that heuristic responding on the scale could be a better indicator of (a lack of) cognitive reflection in populations with lower levels of the cognitive reflection trait than correct scores.

We also explored developmental changes in cognitive reflection by comparing correct and heuristic responding across the two age groups. As expected, participants in the younger group produced more heuristic, and fewer correct responses. This was the case for all items, apart from item 2 (an item from the original scale), where we found no developmental difference in either heuristic or correct responding. An additional analysis also confirmed that the new items discriminated better

between the two age groups than the old items. This also implies that, when the scale is administered to adult participants, adding the new items can help to discriminate more precisely between respondents with lower to medium levels of the cognitive reflection trait.

Although the present results are in line with earlier studies which investigated developmental changes in heuristic reasoning between adolescence and adulthood (e.g., Chiesi et al., 2011; Klaczynski, 2001a; Morsanyi & Handley, 2013), it should be noted that the groups not only differed in their age. Indeed, it is very likely that university students had higher levels of numeracy, and they obviously represent a population with above-average levels of intelligence and motivation, which could also explain the differences in cognitive reflection between the two age groups. Nevertheless, regardless of whether these differences reflect developmental changes or other differences between the two samples, the present results clearly demonstrate the suitability of the CRT-Long to be used with populations with relatively low levels of cognitive reflection.

General discussion

Since its publication in 2005, Frederick's paper on the CRT has been cited over 700 times. Indeed, the CRT is an extremely popular measure, as it only takes a few minutes to administer, and, at the same time, it is one of the best predictors of rational thinking and decision-making skills. Nevertheless, despite the widespread use of the CRT, its psychometric properties have not been extensively investigated. In his original publication, Frederick (2005) did not report the reliability of the scale, and, with a few exceptions (Campitelli & Gerrans, 2013; Liberali et al., 2012; Morsanyi et al., 2014; Weller et al., 2012) most researchers who used the scale followed the same practice.

In the present studies, for the first time we aimed at investigating the psychometric properties of the CRT by employing Item Response Theory (IRT) in which characteristics of items in a test (i.e., item parameters) and the characteristic of individuals (i.e., a latent trait) are related to the probability of a positive response (i.e., a trait-consistent endorsement of an item). Applying IRT, we measured the

CRT items' psychometric properties. Preliminarily, we tested the unidimensionality of the original CRT, which was never tested before. This is a desirable characteristic, as a single-factor structure facilitates the scale's function as a population screen of the cognitive reflection trait.

Our analyses demonstrated that the original CRT items had high discriminative power i.e., they were able to distinguish between respondents with different levels of the cognitive reflection trait around their location, which was above the mean. Given the difficulty of the original items, a large proportion of educated adults score zero on the scale. For this reason, we developed a longer version of the scale with the aim of measuring a wider range of the trait. The CRT-L also showed unidimensionality, and the properties of the new items (i.e., difficulty and discrimination) were consistent with our aims. With regard to difficulty, one of the new items was located around the mean, one below the mean, and one above the mean. Concerning discrimination, the parameter estimates indicated that the items of the CRT-L were able to distinguish well between different levels of the trait.

Additionally, our validity results clearly showed that the CRT-L (similarly to the CRT) had a significant mathematics component (see also Campitelli & Gerrans, 2014; Libearli et al., 2011; Welsh et al., 2013) but it was also related to decision-making skills (Frederick, 2005), measures of rational thinking and reasoning ability (Toplak et al., 2011; 2014), and thinking dispositions (e.g., Fernbach et al., 2013; Mata et al., 2013; Pennycook et al., 2012). Overall, these results confirmed that the CRT-L, just like the original CRT, is a reliable and powerful predictor of a person's ability to make unbiased judgments and rational decisions in a wide variety of contexts.

Given that the CRT is related to a wide range of measures (including measures of numeracy, intelligence, thinking dispositions, and reasoning and decision making skills), the question arises what cognitive reflection really is, and whether it exists as an independent construct. For example, would it be advantageous to develop a measure of cognitive reflection which shows weaker correlations with measures of numeracy or intelligence? Frederick (2005, *p*. 26) described the CRT as "a simple measure

of one type of cognitive ability", as well as "the ability or disposition to resist reporting the response that first comes to mind" (Frederick, 2005, *p.* 35). Some other researchers described the CRT as an ability measure, whereas others consider it both as an ability and a disposition measure (see Campitelli & Gerrans, 2013 for a review). Although this question is theoretically important, the aim of the present studies was not to better understand the construct of cognitive reflection or to develop a version of the scale which is less strongly related to numeracy, intelligence or thinking dispositions. Indeed, we expect that in that case the CRT would lose some of its predictive power. Rather we aimed at developing a version which preserves the original properties of the CRT in terms of its relationships with other constructs, but it is also appropriate for participants with lower levels of the trait. Thus, similarly to the original items, all new items included numerical content, and predominantly elicited either a correct response or a typical incorrect (i.e., heuristic) response.

Similarly to previous studies (e.g., Albaity, Rahman & Shahidul, 2014; Campitelli & Gerrans, 2013; Frederick, 2005; Toplak et al., 2014), we found a gender difference on the CRT, as well as on the CRT-L. Our analyses also suggested that the numerical content of the problems might be responsible for these gender differences. Although designing problems with no numerical content might eliminate the gender differences, without the numerical content it might be difficult to design problems where there are clear normative standards (i.e., where we can confidently say that the typical heuristic response is incorrect). Indeed, a core idea behind the concept of cognitive reflection is that the effortless, heuristic responses are incorrect, and correct responses can only be generated through effortful processing.

Overall the results attested that the CRT-L scale is more adequate for younger and less educated samples than the original CRT. This is a very important improvement over the original CRT, as it makes it possible to investigate changes in cognitive reflection across the life span. A related potential direction of future research is to examine whether cognitive reflection shows similar relationships with

measures of decision making, reasoning, intelligence, thinking dispositions and numeracy in the case of adolescents and older adults as in the case of highly educated younger adults (the typical target population of cognitive reflection research). This could give further insight into developmental changes in thinking skills and dispositions.

Finally, more items can also be advantageous in investigating the potential effects of experimental manipulations on cognitive reflection. Indeed, such studies could be very informative with regard to the optimal conditions for rational decision making and reasoning, as well as regarding circumstances which are the most likely to lead to incorrect heuristic responses. For example, Alter, Oppenheimer, Epley and Eyre (2007, Experiment 1) showed that students from an elite university were more likely to give correct responses to the CRT when the problems were presented in a hard-to-read font, whereas Morsanyi et al. (2014, Experiment 3) showed that working memory load reduced cognitive reflection.

We acknowledge that the CRT-L might not include a complete range of difficulty, and that our results are limited by the number of items that were included in the initial item pool. In fact, examination of the item parameters would suggest that more items could be added in order to more finely differentiate among respondents, especially between those at the extreme ends of cognitive reflection ability. Future research using IRT analyses can help to create an adaptive test that may assess cognitive reflection across a wider range of the trait. Future work could also address further the question of gender differences in cognitive reflection, and whether it was possible to develop a scale to measure the tendency to give incorrect heuristic responses without the inclusion of numerical content. Nevertheless, the CRT-L offers a useful extension to the original scale, and it should inspire further research into cognitive reflection, including previously neglected topics, such as the issue of developmental changes, and the effect of experimental manipulations.

References

Albaity, M., Rahman, M., & Shahidul, I. (2014). Cognitive reflection test and behavioral biases in Malaysia. *Judgment & Decision Making*, *9*(2).

Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General*, *136*, 569-576.

Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). ERIC Clearinghouse on Assessment and Evaluation. Retrieved from http://info.worldbank.org/etools/

docs/library/117765/Item%20Response%20Theory%20-%20F%20Baker.pdf

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443–458.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). IRTPRO 2.1 for Windows. Chicago, IL: Scientific Software International.

Campitelli, G., & Gerrans, P. (2013). Does the cognitive reflection test measure cognitive reflection? A mathematical modeling approach. *Memory & Cognition*, *42*, 434–447.

Chen, W.H., & Thissen, D. (1997). Local dependence indices for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22, 265-289.

Chiesi, F., Donati, M. A., Papi, C., & Primi, C. (2010). Misurare il pensiero superstizioso nei bambini: validità e attendibilità della Superstitious Thinking Scale [Measuring superstitious thinking in children: validity and reliability of the Superstitious Thinking Scale]. Età Evolutiva, 97, 9–19.

Chiesi, F., Ciancaleoni, M., Galli, S., & Primi, C. (2012). Using the Advanced Progressive Matrices (Set I) to Assess Fluid Ability in a Short Time Frame: An Item Response Theory–Based Analysis. *Psychological Assessment*. DOI:10.1037/a0027830.

Chiesi, F., Primi, C., & Morsanyi, K. (2011). Developmental changes in probabilistic reasoning: The role of cognitive capacity, instructions, thinking styles, and relevant knowledge. *Thinking & Reasoning*, *17*, 315-350.

Cenkseven-Önder , F. (2012). The influence of decision-making styles on early adolescents' life satisfaction. *Social Behavior and Personality*, *40*(9), 1523-1536. http://dx.doi.org/10.2224/sbp.2012.40.9.1523

Cokely, E.TY. & Kelly, C.M. ("009). Cognitive abilities and superior decision making under risk: A protocol analysis and process model evaluation. *Judgment and Decision Making*, 4, 20-33.

De Neys, W., Rossi, S., & Houdé, O. (2013). Bats, balls, and substitution sensitivity: Cognitive misers are no happy fools. *Psychonomic Bulletin & Review*, *20*, 269-273.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Evans, J. S. B., Handley, S. J., Neilens, H., & Over, D. E. (2007). Thinking about conditionals: A study of individual differences. *Memory & cognition*, *35*(7), 1772-1784.

Fagerlin, A., Zikmund-Fisher, B. J., Ubel, P. A., Jankovic, A., Derry, H. A., & Smith, D. M. (2007). Measuring numeracy without a math test: development of the Subjective Numeracy Scale. *Medical Decision Making*, *27*(5), 672-680.

Fernbach, P. M., Sloman, S. A., Louis, R. S., & Shube, J. N. (2013). Explanation fiends and foes: How mechanistic detail determines understanding and preference. *Journal of Consumer Research*, *39*(5), 1115-1131.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19,* 25–42.

Gauffroy, C., & Barrouillet, P. (2009). Heuristic and analytic processes in mental models for conditionals: An integrative developmental theory. *Developmental Review*, *29*, 249-282.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky, *Psychological Review, 103*, 592–596.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hertwig, R., Benz, B., & Krauss, S. (2008). The conjunction fallacy and the many meanings of ‘‘and’’. *Cognition, 108*, 740–753.

Hyde, J.S., Fennema, E., & Lamon, S.J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin, 107*, 139–155.

Janis, I.L., & Mann, L. (1977). Decision-making: A psychological analysis of conflict, choice and commitment. New York: Free Press.

Klaczynski, P. A. (2001a). Framing effects on adolescent task representations, analytic and heuristic processing, and decision making: Implications for the normative-descriptive gap. *Journal of Applied Developmental Psychology, 22*, 289-309.

Klaczynski, P. A. (2001b). The influence of analytic and heuristic processing on adolescent reasoning and decision making. *Child Development, 72*, 844-861.

Klaczynski, P. A., & Cottrell, J. E. (2004). A dual-process approach to cognitive development: The case of children’s understanding of sunk cost decisions. *Thinking & Reasoning, 10*, 147-174.

Kokis, J. V., MacPherson, R., Toplak, M. E., West, R. F., & Stanovich, K. E. (2002). Heuristic and analytic processing: Age trends and associations with cognitive ability and cognitive styles. *Journal of Experimental Child Psychology, 83*, 26–52.

Liberali, J. M., Reyna, V. F., Furlan, S., Stein, L. M. & Pardo, S. T. (2012). Individual differences in numeracy and cognitive reflection, with implications for biases and fallacies in probability judgment. *Journal of Behavioral Decision Making*. DOI: 10.1002/bdm.752

Lipkus, I. M., Samsa, G. & Rimer, B.K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making*, 21, 37-44.

Marcoulides, GA, & Hershberger, SL. (1997). *Multivariate statistical methods*. A first course. Mahawa, NJ: Lawrence Erlbaum Associates)

Mata, A., Ferreira, M. B., & Sherman, S. J. (2013). The metacognitive advantage of deliberative thinkers: A dual-process perspective on overconfidence. *Journal of Personality and Social Psychology, 105,* 353–373.

Morsanyi, K., Busdraghi, C., & Primi, C. (2014). Mathematical anxiety is linked to reduced cognitive reflection: a potential road from discomfort in the mathematics classroom to susceptibility to biases. *Behavioral and Brain Functions*, 10:31. doi:10.1186/1744-9081-10-31 1254

Morsanyi, K., Devine, A., Nobes, A., & Szűcs, D. (2013). The link between logic, mathematics and imagination: evidence from children with developmental dyscalculia and mathematically gifted children. *Developmental Science, 16,* 542-553.

Morsanyi, K., & Handley, S.J. (2013). Heuristics and biases – Insights from developmental studies. In: P. Barouillet & C. Gauffroy (Eds.) *The development of thinking and reasoning* (pp. 122-149) Hove: Psychology Press.

Muthén, L. K., & Muthén, B. O. (2004). *Mplus: The comprehensive modeling program for applied researchers. User's guide* (3rd ed.). LosAngeles, CA: Muthén & Muthén**.**

Pennycook, G., Cheyne, J. A., Seli, P., Koehler, D. J., & Fugelsang, J. A. (2012). Analytic cognitive style predicts religious and paranormal belief. *Cognition, 123*, 335–346.

Primi, C., Morsanyi, K., & Chiesi, F. (2014). Measuring the basics of probabilistic reasoning: The IRT-based construction of the probabilistic reasoning questionnaire. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg,

The Netherlands: International Statistical Institute. iase-web.org [© 2014 ISI/IASE]

Raven, J. C. (1941). Standardization of progressive matrices. *British Journal of Medical Psychology, 19,* 137–150.

Raven, J. C. (1962). Advanced progressive matrices. London: Lewis & Co. Ltd.

Rust, J., & Golombok, S. (Eds.). (1999). *Modern psychometrics: The science of psychological assessment* (2nd ed.). New York: Routledge.

Stone, C. A., & Zhang, B. (2003). Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement, 40,* 331–352.

Thompson, V.A. & Morsanyi, K. (2012). Analytic thinking: Do you feel like it? *Mind & Society, 11,* 93-105.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics and biases tasks. *Memory & Cognition, 39,* 1275-1289.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2013). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning,* DOI: 10.1080/13546783.2013.845605.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211,* 453-458.

Van Dooren, W., De Bock, D., Hessels, A., Janssens, D., Verschaffel, L. (2005). Remedying secondary school students' illusion of linearity: developing and evaluating a powerful learning environment. In: Verschaffel L., e.a. (Eds.), Powerful environments for promoting deep conceptual and strategic learning (Studia paedagogica, 41) (pp. 115-132). Leuven: Universitaire Pers.

Weller, J. A., Dieckmann, N. F., Tusler, M., Mertz, C. K., Burns, W. J., & Peters, E. (2013). Development and testing of an abbreviated numeracy scale: A rasch analysis approach. *Journal of Behavioral Decision Making, 26*(2), 198-212.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson tests of achievement*. Itasca, IL: Riverside Publishing.

Table 1. *Percentage of correct and heuristic responses, standardized factor loadings, fit statistics, and parameters for each item of the Cognitive Reflection Test (CRT) based on Study 1.*

Item	*% C (H)*	*λ*	*S-χ^2(df)*	*b (SE)*	*a (SE)*
1	43 (56)	.70	.05 (1)	.26 (.09)	1.70 (.32)
2	34 (57)	.66	.04 (1)	.70 (.12)	1.48 (.26)
3	47 (46)	.90	.03 (1)	.08 (.07)	3.05 (.75)

Note. % represents the percentage of correct (C) and heuristic (H) responses. Standardized factor loadings λ are all significant at $p = .001$. The parameters were computed under the 2PL model (a = discrimination, b = difficulty).

Table 2. *Percentages of correct and heuristic responses, standardized factor loadings, fit statistics, and parameters for each item of the CRT and CRT-L based on Study 2.*

Item CRT	*% C (H)*	*λ*	*S-χ2(df)*	*b (SE)*	*a (SE)*
1	*39 (49)*	*.65*	*.65 (1)*	*.43 (.07)*	*1.45 (.16)*
2	*44 (48)*	*.74*	*.61 (1)*	*.21 (.06)*	*1.90 (.24)*
3	*54 (36)*	*.89*	*.70 (1)*	*-.12 (.07)*	*3.26 (.68)*
Item CRT-L	*% C (H)*	*λ*	*S-χ2(df)*	*b (SE)*	*a (SE)*
1	39 (49)	.70	6.30 (4)	.39 (.06)	1.69 (.16)
2	44 (48)	.74	7.34 (4)	.20 (.05)	1.89 (.18)
3	54 (36)	.85	13.07 (4)	-.15 (.05)	2.73 (.29)
4	81 (15)	.77	9.42 (4)	-1.19 (.09)	2.03 (.23)
5	37 (36)	.74	7.24 (4)	.42 (.06)	1.86 (.18)
6	49 (37)	.73	1.19 (4)	.01 (.05)	1.79 (.17)

Note. % represents the percentage of correct (C) and heuristic (H) responses. Standardized factor loadings λ are all significant at $p = .001$. Parameters were computed under the 2PL model (a = discrimination, b = severity). Due to the large sample size ($N = 988$), α was fixed at .01.

Table 3. *Descriptive statistics for the individual differences measures, and correlations between the CRT, the CRT-L new items, the CRT-L and all the other variables in the study.*

	M (SD) *Range*	CRT	CRT-Long new items	CRT-Long
APM	8.43 (1.99)	.32***	.31***	.39***
	1-12	(N=201)	(N=201)	(N=201)
PRQ	14.96 (1.44)	.31***	.29***	.33***
	5-16	(N=234)	(N=234)	(N=234)
NS	9.16 (1.72)	.41***	.36***	.44***
	3-11	(N=201)	(N=201)	(N=201)
Math Fluency	109.95 (21.95)	.28*	.42**	.42**
	67-158	(N=59)	(N=59)	(N=59)
SNS	37.70 (5.51)	.19**	.15*	.19**
	21-48	(N=234)	(N=234)	(N=234)
Conditional Probability	3.66 (.67)	.22**	.16**	.22**
	1-4	(N=226)	(N=226)	(N=226)
Conditional Reasoning	6.12 (2.27)	.35**	.29**	.38*
	3-12	(N=59)	(N=59)	(N=59)
Transitive Inferences	5.41 (1.05)	.24□	.30*	.32*
	1-6	(N=59)	(N=59)	(N=59)
Risk Seeking Behaviour	2.74 (1.66)	.18*	.25**	.26**
	0-8	(N=199)	(N=199)	(N=199)
Intertemporal Behaviour	2.20 (1.29)	.13**	.20**	.16**
	0-5	(N=199)	(N=199)	(N=199)
STS	17.55 (6.46)	-.24***	-.23***	-.27***
	8-36	(N= 393)	(N=393)	(N=393)

APM = *Advanced Progressive Matrices*, PRQ= *Probabilistic Reasoning Questionnaire*, NS = Numeracy Scale, SNS = *Subjective Numeracy Scale*, STS = *Superstitious Thinking Scale*.□ p=.067; * p<.05; ** p<.01; *** p<.001

Table 4. *Multiple regression with risk seeking behavior total score as dependent variable and numeracy, fluid intelligence and CRT-L as independent variables.*

Predictors	*B*	*β*	*t*	*p*
Numeracy	-.11	-.09	-1.07	.29
Fluid intelligence	.03	.04	.42	.68
CRT-L	.28	.29	3.23	.002

$F(3,158)=4.33$, $p=.006$; $R=.28$; $R^2=.08$

Table 5. *Multiple regression with probabilistic reasoning score as dependent variable and math fluency, subjective numeracy and CRT-L scores as independent variables.*

Predictors	*B*	*β*	*t*	*p*
Math fluency	.02	.16	1.14	.26
Subjective numeracy	.04	.13	.96	.34
CRT-L	.53	.30	2.10	.04

$F(3,58)=5.23$, $p=.003$; $R=.47$; $R^2=.22$

Table 6. *The proportion of correct and heuristic responses given to each item and the overall proportion of correct and heuristic responses on the CRT and CRT-L across the adolescent and adult samples.*

	Correct responses		Heuristic responses	
	adolescents	adults	adolescents	adults
Item 1	17%	48%	80%	48%
Item 2	27%	31%	62%	62%
Item 3	17%	50%	76%	44%
Item 4	40%	81%	57%	18%
Item 5	17%	39%	61%	49%
Item 6	16%	84%	66%	8%
CRT	20%	43%	72%	51%
CRT-Long	22%	56%	67%	38%

Figure captions

Figure 1. Exemplar Theoretical Item Characteristic Curve for the Two-Parameter Logistic Item Response Theory Model (a = discrimination, b = difficulty).

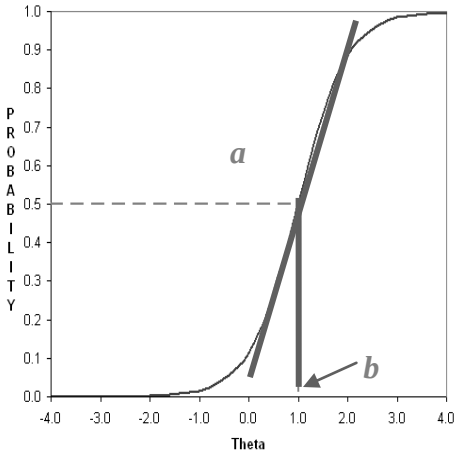
Figure 2. The ICCs of the original Cognitive Reflection Test (CRT) items under the 2PL model in Study 1. The latent trait (Theta) is represented on the horizontal axis and the probability of correct responding is shown on the vertical axis.

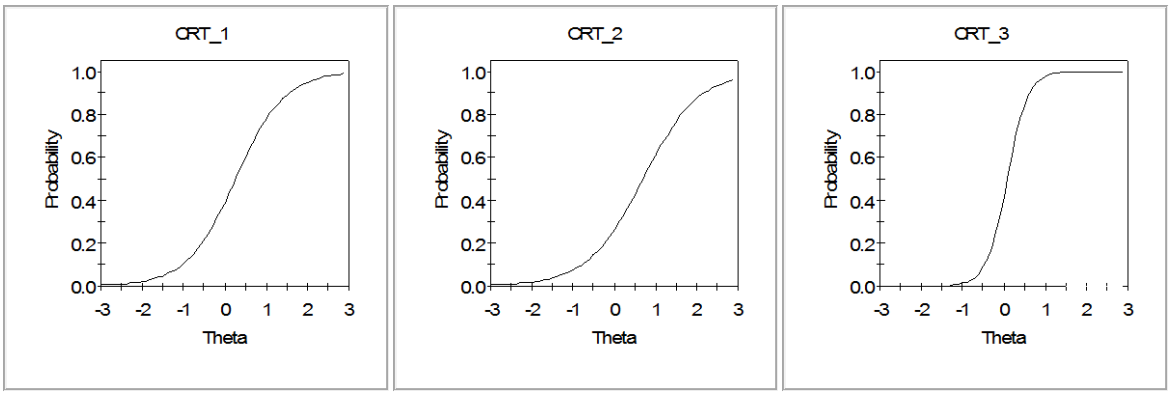
Figure 3. Test information function of the CRT under the 2PL model. Latent trait (Theta) is shown on the horizontal axis, and the amount of information and the standard error yielded by the test at any trait level are shown on the vertical axis.

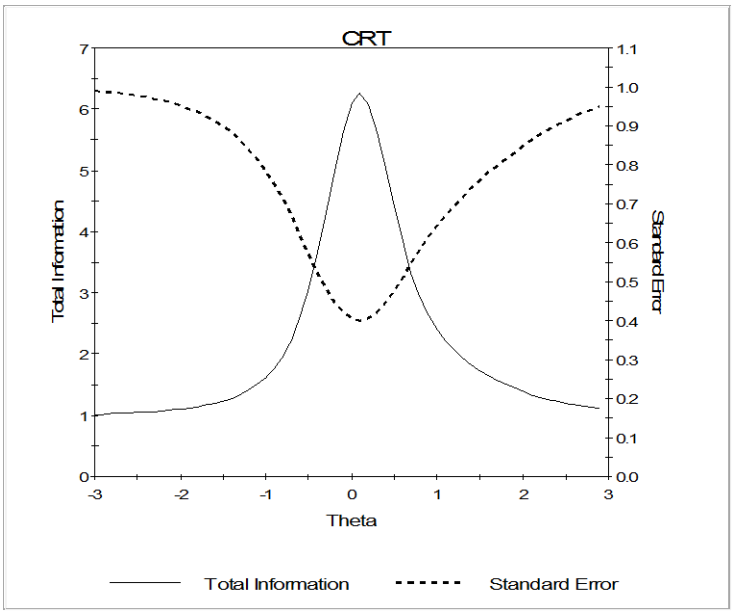
Figure 4. The ICCs of the Cognitive Reflection Test-Long (CRT-L) items under the 2PL model in Study 2. The latent trait (Theta) is represented on the horizontal axis and the probability of correct responding is shown on the vertical axis.

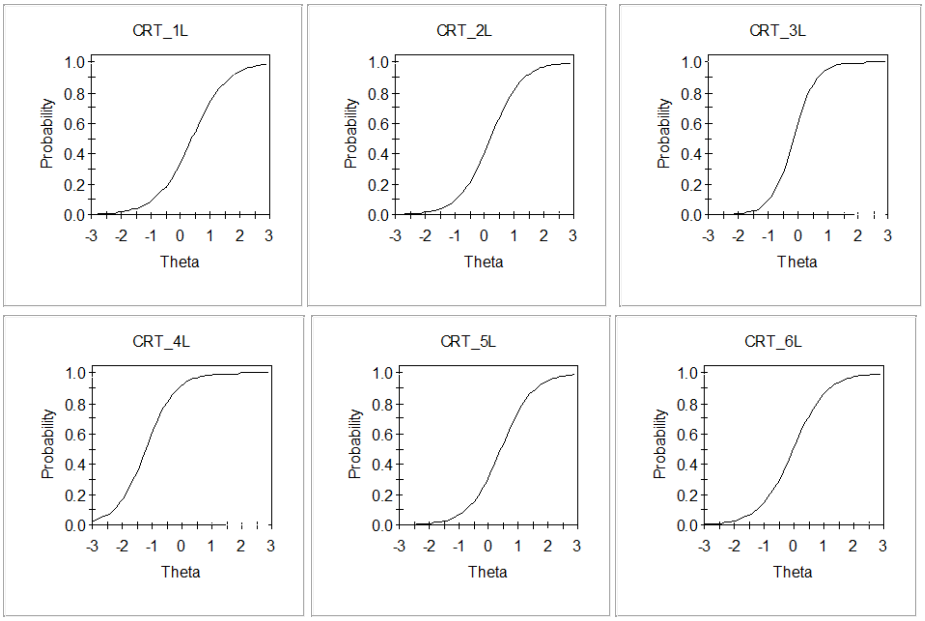
Figure 5. Test information function of the CRT-L under the 2PL model. The latent trait (Theta) is shown on the horizontal axis, and the amount of information and the standard error yielded by the test at any trait level are shown on the vertical axis.

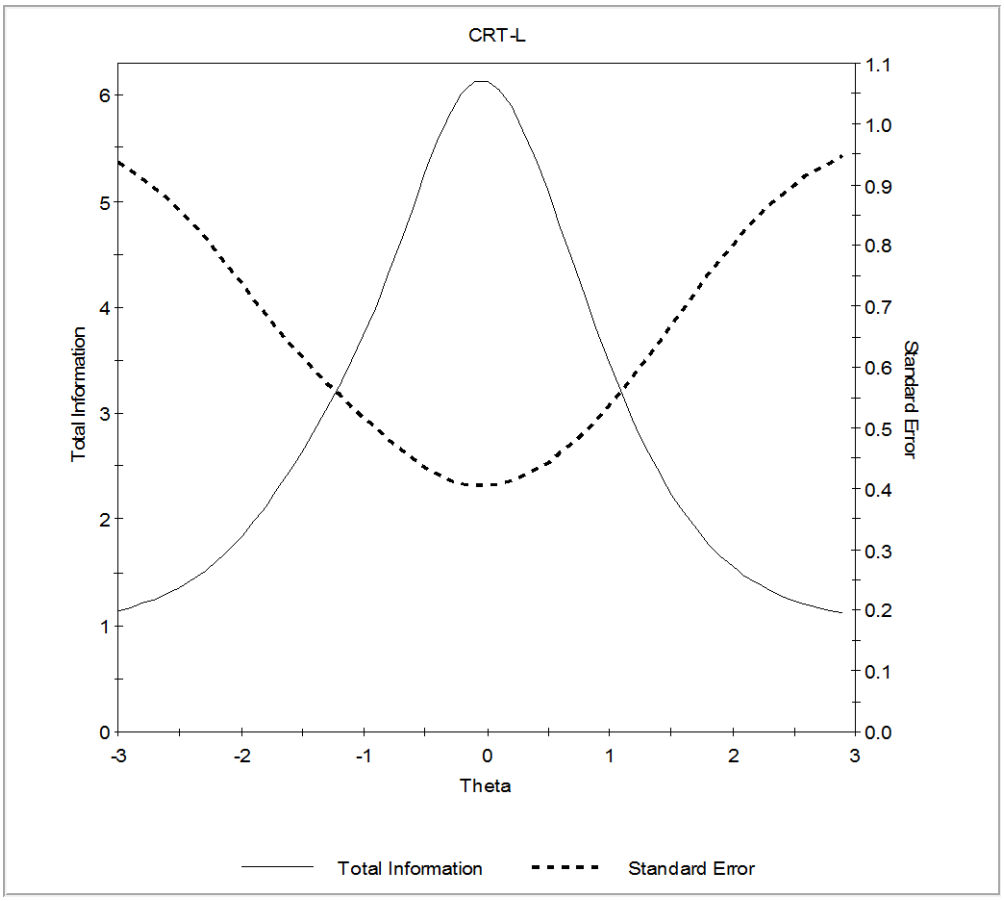
Figure 6. Histograms presenting the distribution of correct and heuristic scores on the CRT-Long for the early adolescents and young adults.

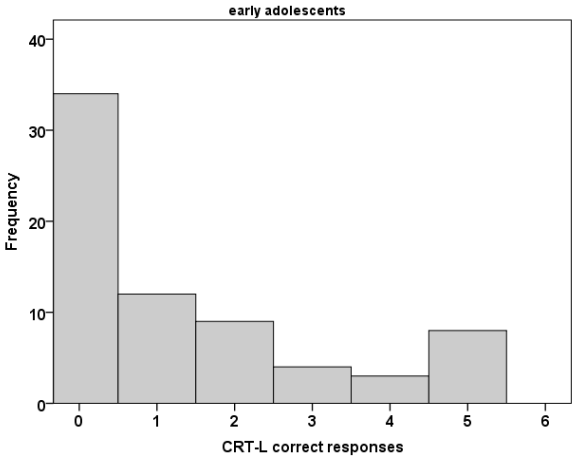

1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
PROBABILITY
a
b
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0
Theta

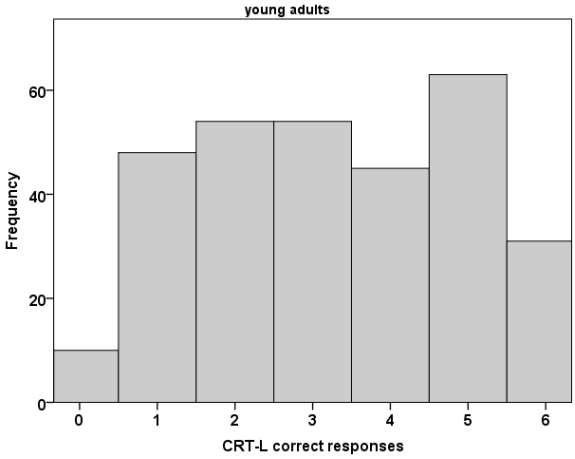

CRT_1
Probability
1.0
0.8
0.6
0.4
0.2
0.0
-3 -2 -1 0 1 2 3
Theta
CRT_2
Probability
1.0
0.8
0.6
0.4
0.2
0.0
-3 -2 -1 0 1 2 3
Theta
CRT_3
Probability
1.0
0.8
0.6
0.4
0.2
0.0
-3 -2 -1 0 1 2 3
Theta

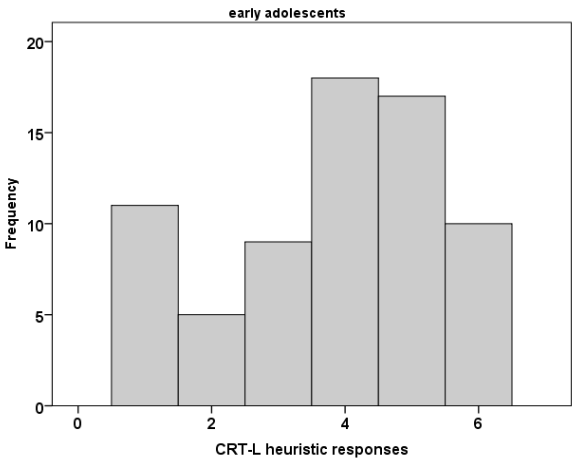

CRT
Total Information
Standard Error
Theta
Total Information
Standard Error

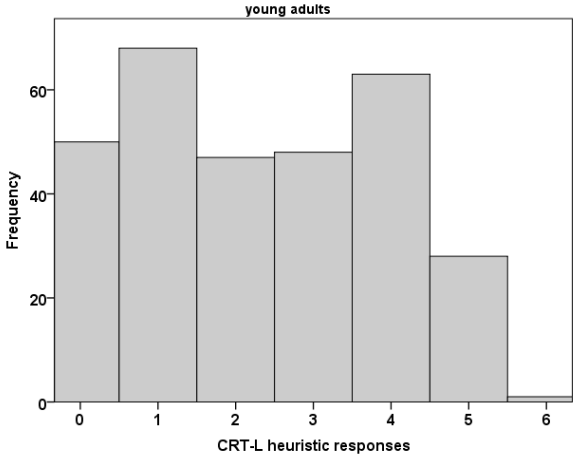

CRT_1L
CRT_2L
CRT_3L
CRT_4L
CRT_5L
CRT_6L
Probability
Theta
1.0
0.8
0.6
0.4
0.2
0.0
-3 -2 -1 0 1 2 3


CRT-L
Total Information
Standard Error
Theta
Total Information
Standard Error
0
1
2
3
4
5
6
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
-3
-2
-1
0
1
2
3


early adolescents
Frequency
CRT-L correct responses


young adults
Frequency
CRT-L correct responses


early adolescents
Frequency
CRT-L heuristic responses


young adults
Frequency
CRT-L heuristic responses

APPENDIX A

Cognitive Reflection Test (CRT)

1. A bat and a ball cost £1.10 in total. The bat costs £1.00 more than the ball. How much does the ball cost? [Correct answer= 5 cents; Heuristic answer= 10 cents]

2. If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets? [Correct answer= 5 minutes; Heuristic answer= 100 minutes]

3. In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake?

[Correct answer= 47 days; Heuristic answer= 24 days]

Additional items included in the CRT-Long

4. If 3 elves can wrap 3 toys in 1 hour, how many elves are needed to wrap 6 toys in 2 hours?

[Correct answer= 3 elves; Heuristic answer= 6 elves]

5. Jerry received both the 15th highest and the 15th lowest mark in the class. How many students are there in the class?

[Correct answer= 29 students; Heuristic answer= 30 students]

6. In an athletics team tall members are three times as likely to win a medal than short members. This year the team has won 60 medals so far. How many of these have been won by short athletes?

[Correct answer= 15 medals; Heuristic answer= 20 medals]

APPENDIX B

Additional items considered during the development of the CRT-L scale

	Items considered during the scale construction process	Step 1	Step 2	Step 3
1*	If you flipped a fair coin 3 times, what is the probability that it would land "Heads" at least once? _____ percent	✓	✓	
2*	A car and a bus are on a collision course, driving toward each other. The car is going 70 miles an hour. The bus is going 80 miles an hour. How far apart are they one minute before they collide?_____ miles	✓		
3*	If John can drink one barrel of water in 6 days, and Mary can drink one barrel of water in 12 days, how long would it take them to drink one barrel of water together? _____ days	✓		
4*	If three elves can wrap six toys in half an hour, how many elves are needed to wrap twenty toys in one hour?_____ elves	✓	✓	✓
5*	Jerry received both the 15th highest and the 15th lowest mark in the class. How many students are in the class?_____ students	✓	✓	✓
6□	Ellen and Kim are running around a track. They run equally fast but Ellen started later. When Ellen has run 5 laps, Kim has run 15 laps. When Ellen has run 30 laps, how many has Kim run? _____laps	✓	✓	
7	An ice cream vendor sells 2/3 of her stock of ice creams on sunny days, and 1/3 of her stock on cloudy days. Yesterday it was a sunny day, and she sold 300 ice creams. Today is a cloudy day. How many can she expect to sell?		✓	
8	In a class there are 42 children. There are 12 more girls than boys. How many girls are there in the class?		✓	
9	In an athletics team tall members are three times as likely to win a medal than short members. This year the team has won 60 medals so far. How many of these have been won by short athletes?		✓	✓

*Items based on Frederick's longer CRT scale

□ Item based on Van Dooren, De Bock, Hessels, Janssens and Verschaffel (2005)